Intelligent Archives in the Context of Knowledge Building Systems - Concepts for the Future

Presented to Intelligent Systems PI Workshop

By

H. K. (Rama) Ramapriyan

Study team

GSFC: Gail McConaughy, Chris Lynnes, Steve Kempler, Ken McDonald

February 4-6, 2004

Rama.Ramapriyan@nasa.gov

Outline

- **Goal and Objectives**
- Assessment of Current State
- Desired State (Phase 1 Findings)
  - Sample scenarios
  - Characteristics
  - Architectural Analysis
- Plans (Phase 2 Approach)
- Conclusion

Goal

To create a next generation conceptual archive architecture supported by advanced technology that is able to:

- Increase data utilization by hosting and applying IDU technologies such as:
  - Information and knowledge extraction
  - Automated data object identification and classification
  - Intelligent system management
  - Distributed computing and data storage
- Automate the transformation of data to information and knowledge allowing the user to focus on research/applications rather than data and data system manipulation
- Exploit new and emerging technologies as they become available
- Incorporate lessons learned from existing archives
- Accommodate new data intensive missions without redesign or restructuring

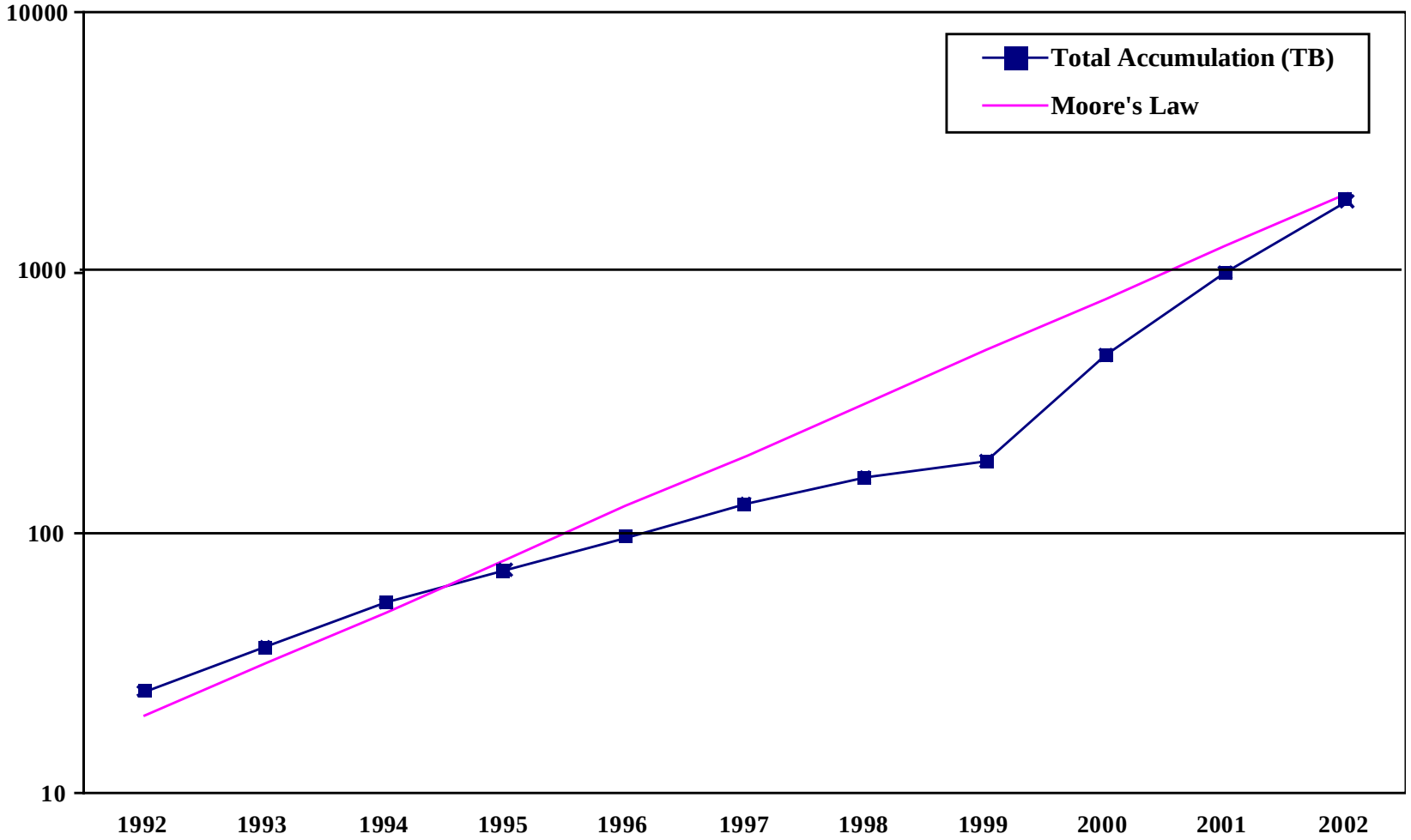
Technical Objectives

- Formulate concepts and architectures that support data archiving for NASA science research and applications in the 10 to 20 year time frame
- Focus on architectural strategies that will support intelligent processes and functions
- Identify and characterize science and applications scenarios that drive intelligent archive requirements
- Assess technologies and research that will need the development of an intelligent archive
- Identify and characterize potential research projects that will be needed to develop and create an intelligent archive

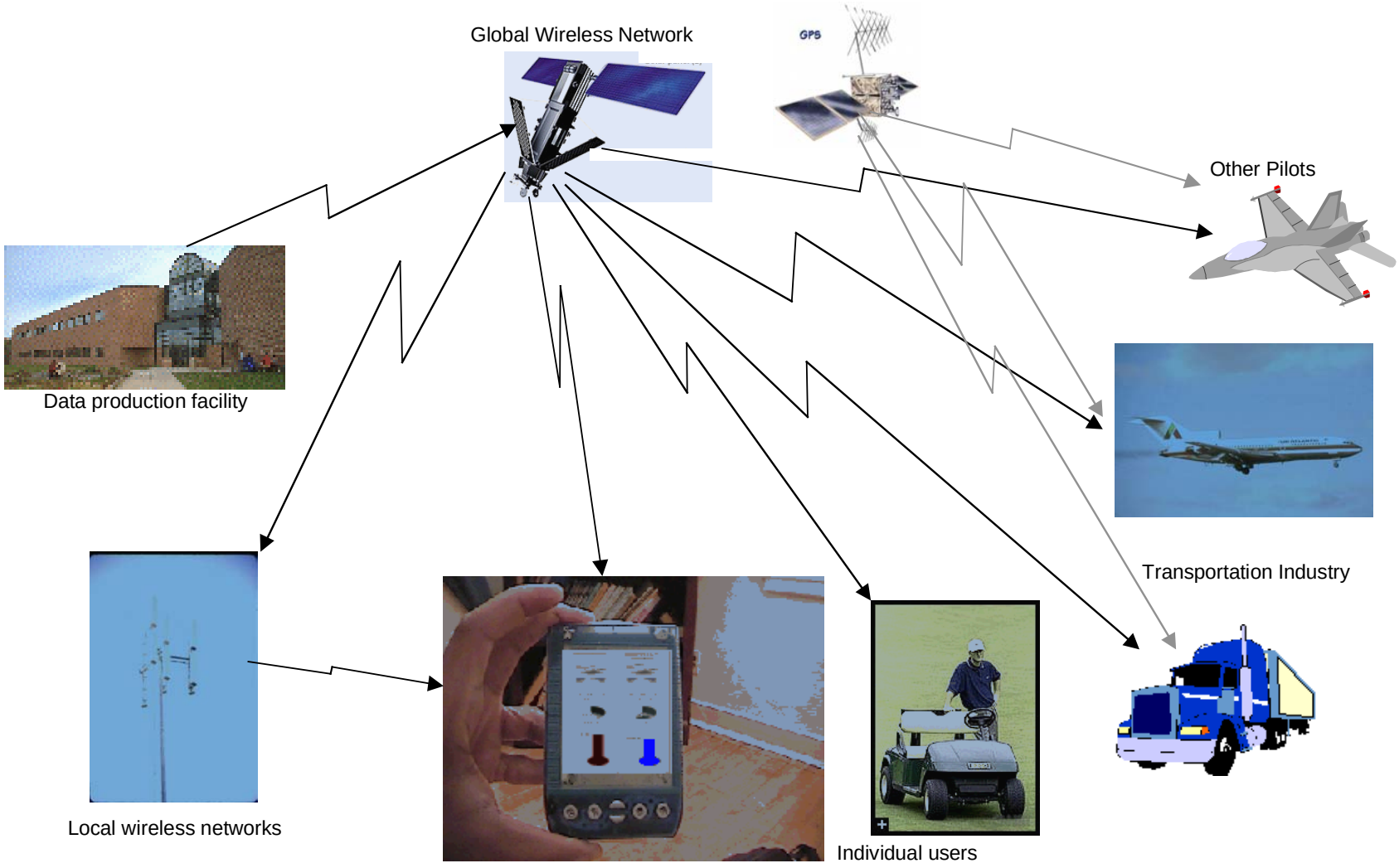
Outline

- Goal and Objectives
- **Assessment of Current State**
- Desired State (Phase 1 Findings)
  - Sample scenarios
  - Characteristics
  - Architectural Analysis
- Plans (Phase 2 Approach)
- Conclusion

Earth Science Data Archive Volume Growth and Moore's Law



Distributed Environment - sensors, providers, users



ESE Data Center Locations

- A total of 50 widely distributed data centers (some of which are at the same location).
- ESE has recently updated our peer reviewed data and information producing centers through the Research, Education and Applications, Solutions Network Cooperative Agreement Notice (REASoN CAN) for development of next-generation architectures.

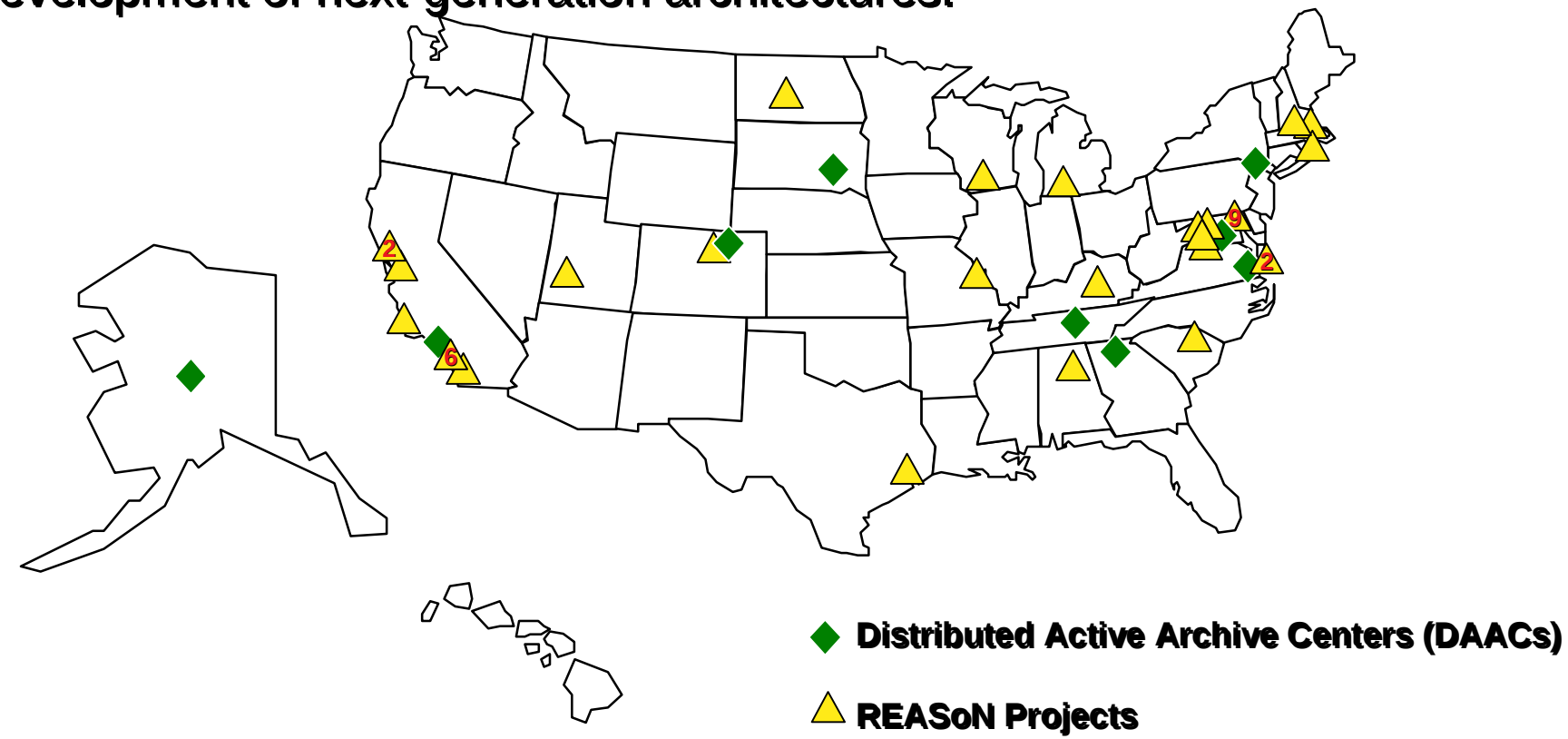


Chart from Frank Lindsay/Martha Maiden – NASA HQ

Assessment of Current State

Data Access

- Wide variety of interfaces and interoperability
- Special services via “focused” research and applications data providers

Timeliness of Data

- Pipeline processes automated, but error recovery tends to be manual
- A few prototype applications of near-real time data uses

Readiness of Data for Use

- Multiple formats and user-developed tools (e.g. format conversion, some subsetting) are in use
- Not all products integrate readily into Geographic Information Systems (GIS)

Data System Responsiveness

- Tending towards smaller heterogeneous and distributed systems
- Feedback does not exist throughout the end-to-end system (e.g., an applications user cannot request sensors to collect additional data based on analysis results)

Data Understandability

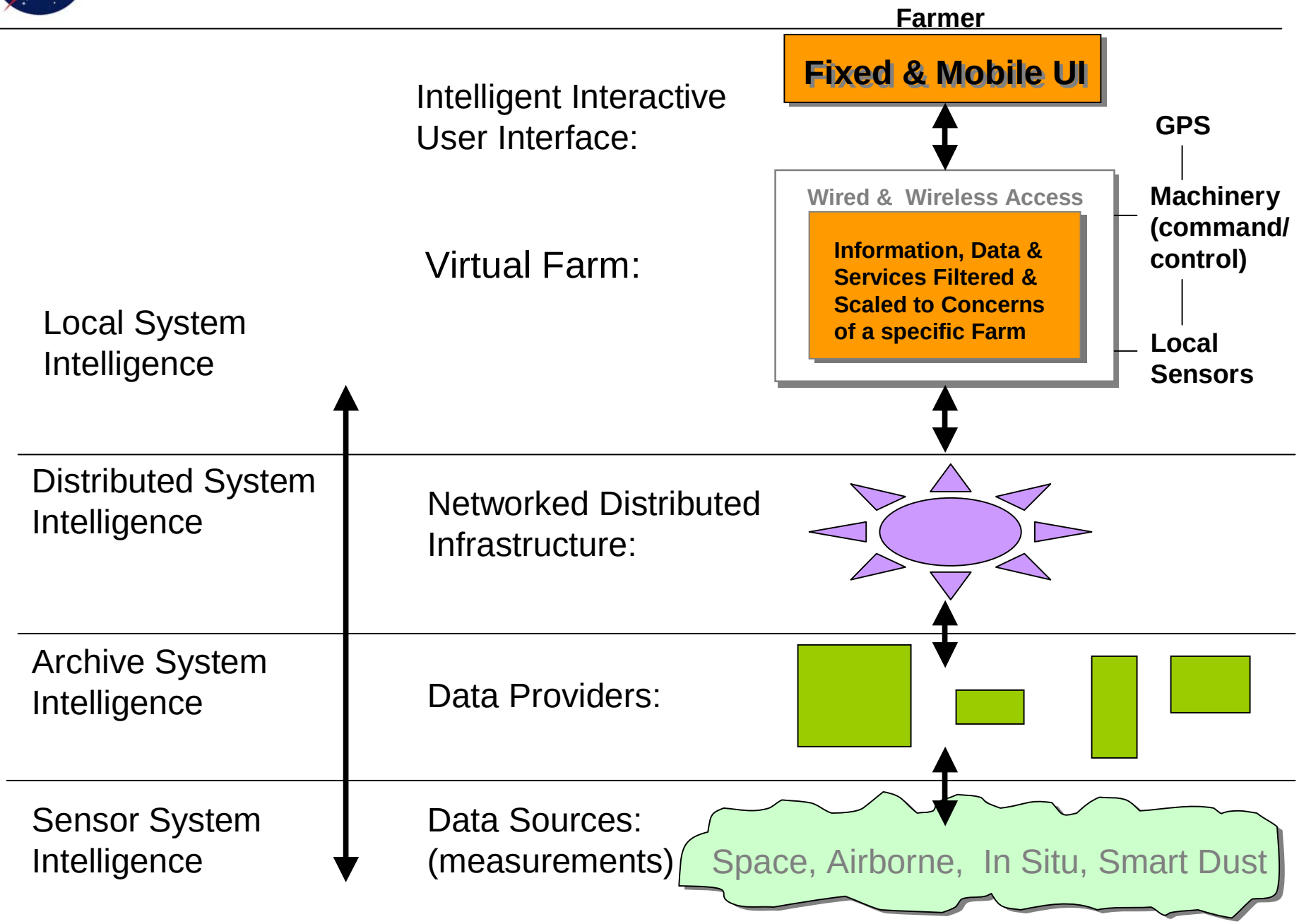
- Quality summaries (ATBDs & Guide documents) and metadata
- Education and Application REASoN Projects (*E* and *A* of REASoN) funded by Code Y

End-to-end significant tailoring still required

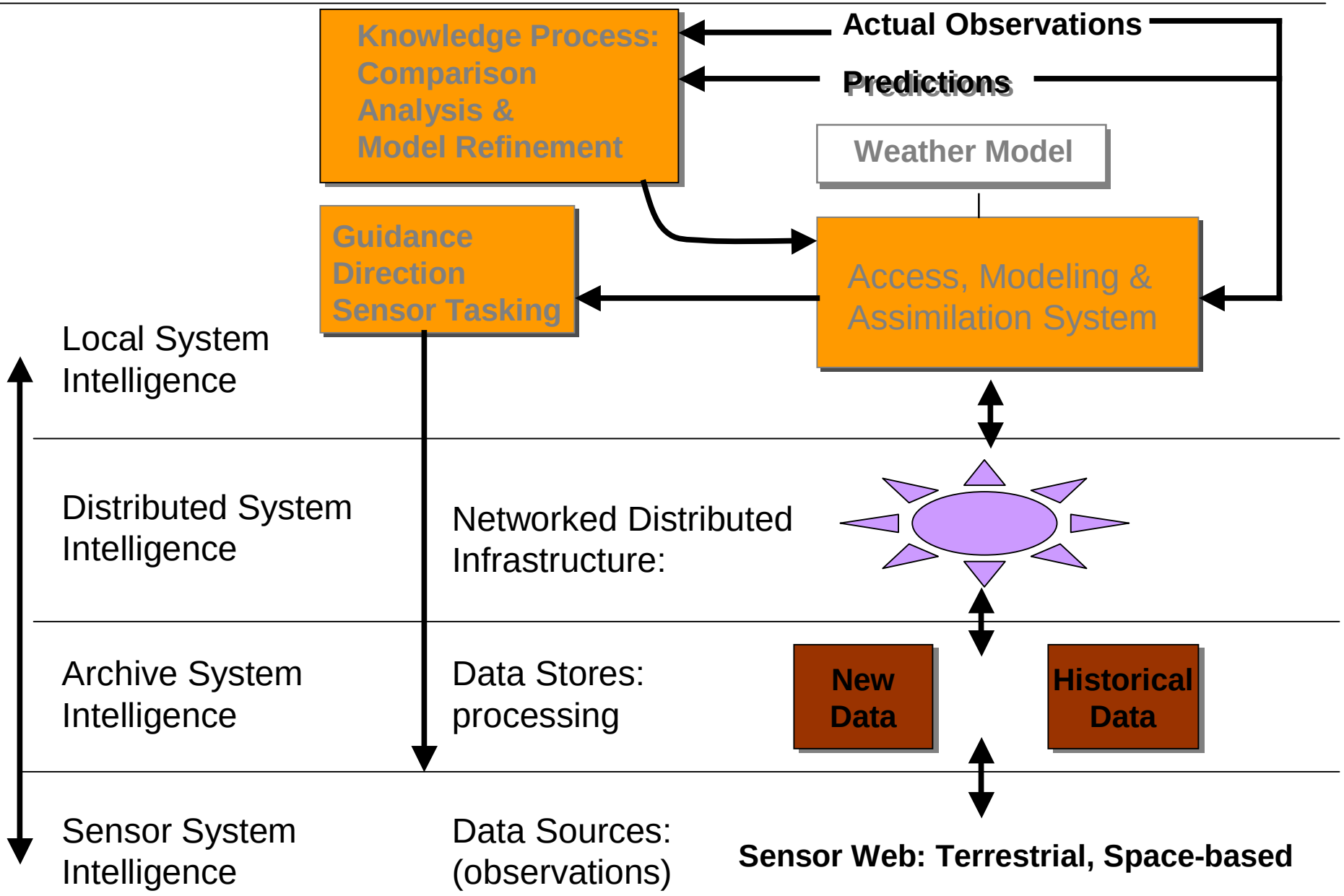
Outline

- Goal and Objectives
- Assessment of Current State
- **Desired State (Phase 1 Findings)**
  - **Sample scenarios**
  - Characteristics
  - Architectural Analysis
- Plans (Phase 2 Approach)
- Conclusion

Precision Agriculture Scenario



Advanced Weather Prediction Scenario



Outline

- Goal and Objectives
- Assessment of Current State
- **Desired State (Phase 1 Findings)**
  - Sample scenarios
  - **Characteristics**
  - Architectural Analysis
- Plans (Phase 2 Approach)
- Conclusion

Data Utilization - What's Needed

Timeliness of Data

- Research use: Near real-time availability generally not critical
- Operational use: Generally requires near real-time availability

Data Access

- Automated Data Discovery and Usage via machine-to-machine interfaces
  - Knowing who holds data of interest in a highly distributed environment
  - Locating just (and all of) what is needed
- Appropriate bandwidth for downloading/Quick ways for ordering via media
- Appropriate capacities of systems (providers' and users')

Data Understandability

- Documentation (machine-readable and machine-usable)
- Detailed (e.g., Pixel-level) Quality information
- Embedded knowledge of what particular data are needed for given operational use
- Etc.

Data Utilization - What's Needed

System Autonomy

- Holdings management
- User Services
- System Management

Readiness of Data for Use

- Plug and play reading tools
- Server-side subsetting, subsampling and other data reduction methods
- Near-universal compatibility with GIS clients

System Responsiveness

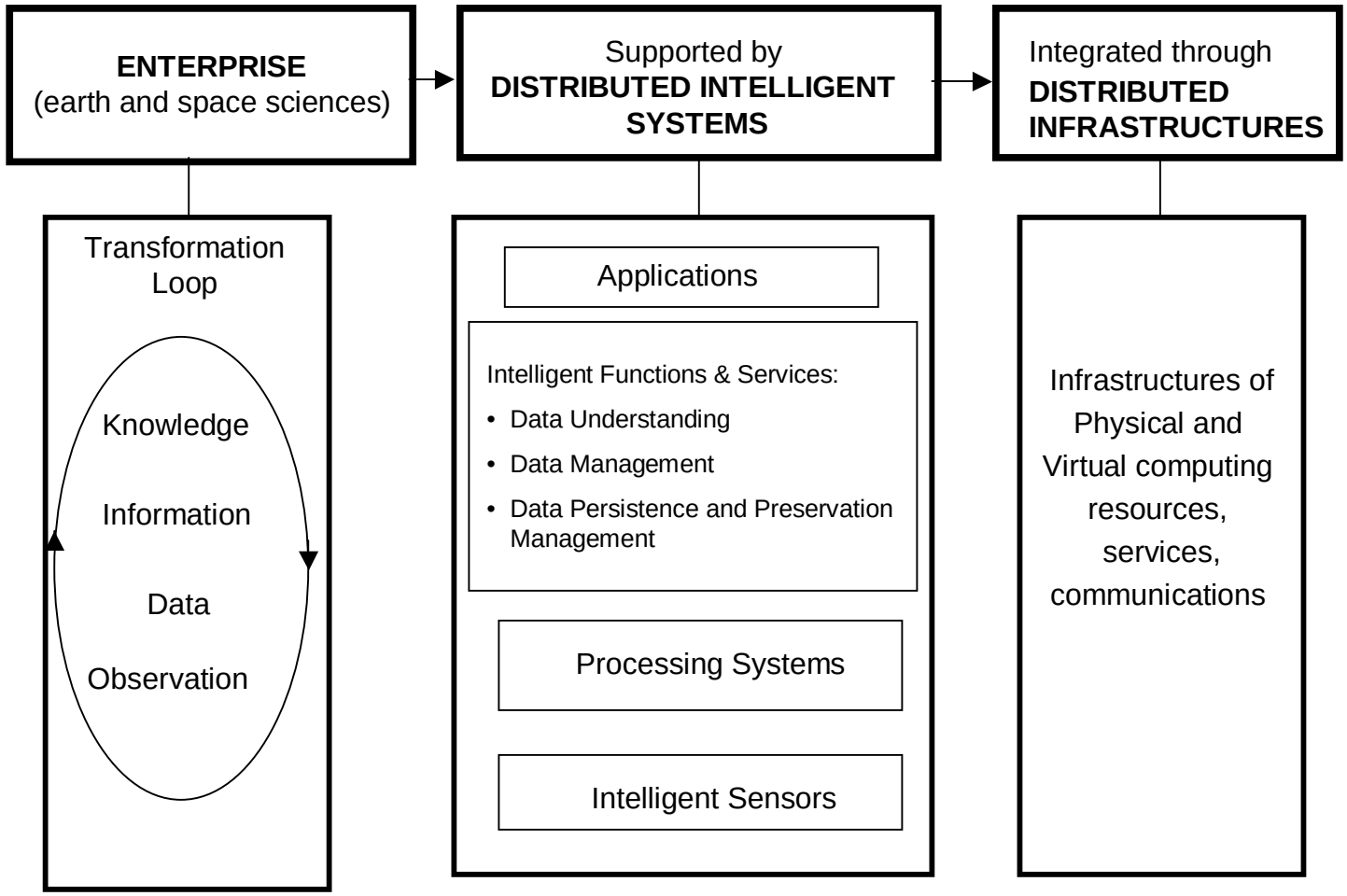
- Ability of "systems" to respond to feedback [note: *"systems"* refers to hardware, software and people]

Most of all: "Effortless" transformation from data to information to knowledge

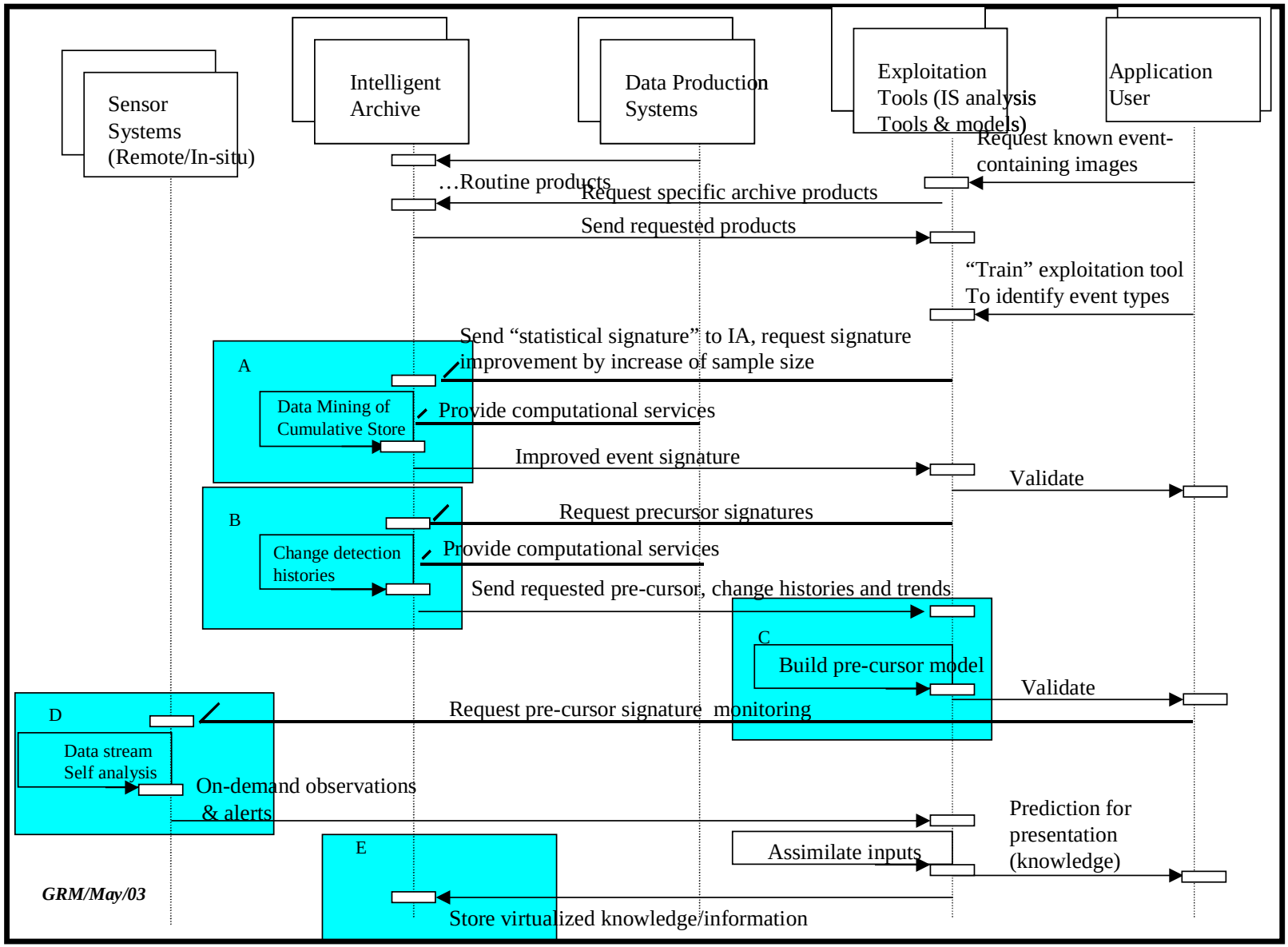
Outline

- Goal and Objectives
- Assessment of Current State
- **Desired State (Phase 1 Findings)**
  - Sample scenarios
  - Characteristics
  - **Architectural Analysis**
- Plans (Phase 2 Approach)
- Conclusion

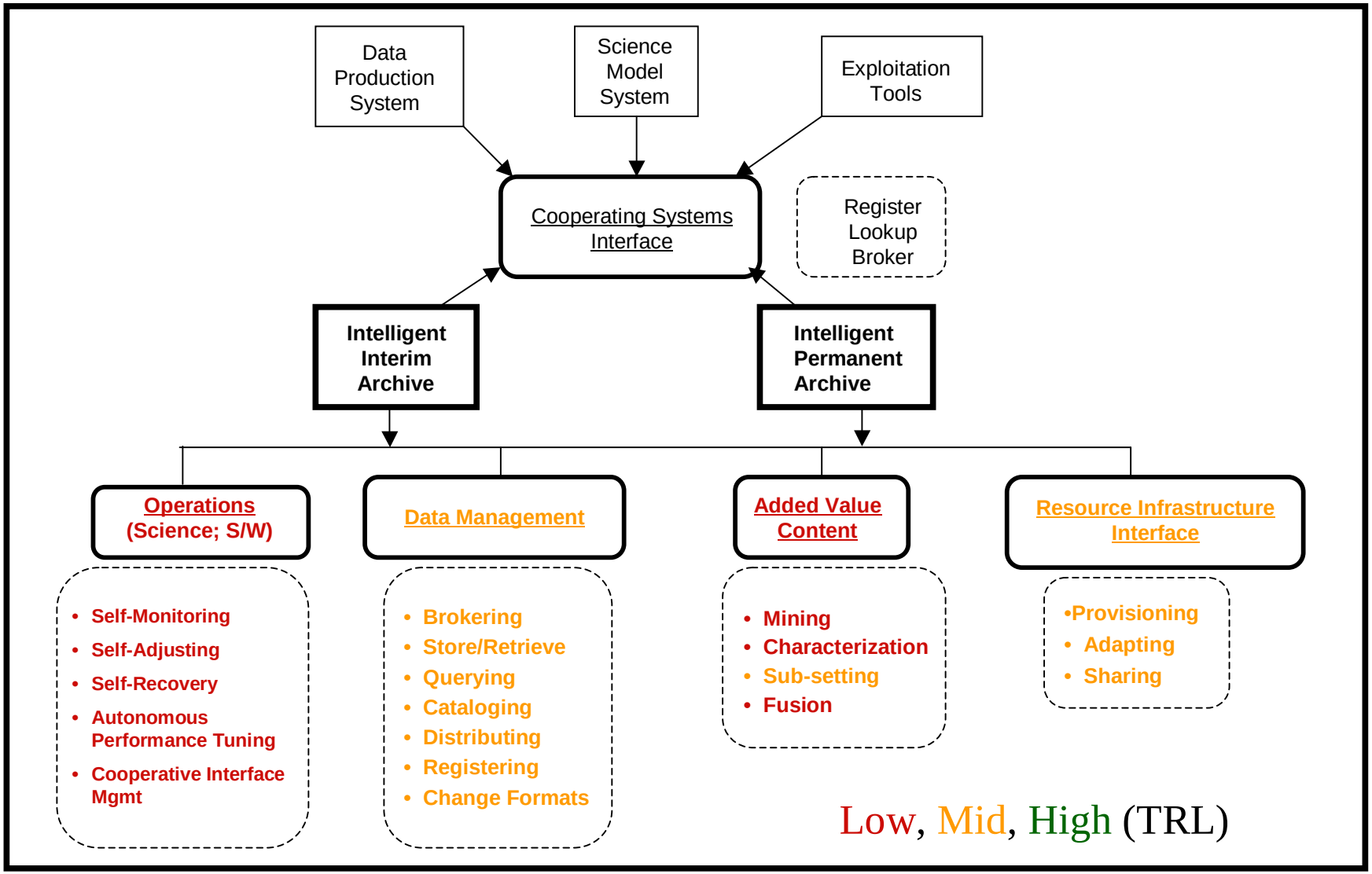
Context - A Knowledge-Building System



A Knowledge-Building Scenario



A Model of IA Focused on Objects and Functions



Outline

- Goal and Objectives
- Assessment of Current State
- Desired State (Phase 1 Findings)
  - Sample scenarios
  - Characteristics
  - Architectural Analysis
- **Plans (Phase 2 Approach)**
- Conclusion

Plans

Examine architectural transformation to knowledge building systems enabled by infusion of IDU/CICT technologies

- Project future NASA mission requirements
  - Identify key problems to be solved by IDU research
  - Quantify scope and scale issues that may prevent such research from eventual utility
- Examine IDU/CICT technologies that will address the issues and have a potential to cope with scope and scale of next 5 to 7 years
  - Define a “specific” design architecture that will demonstrate knowledge building system with representative data volumes, timeliness & usability, merging NASA requirements, IDU/CICT research into a transformative architecture
  - Extend to “speculative architectures” for broader applicability and longer-range transformation 7-10 years

Outline

- Goal and Objectives
- Assessment of Current State
- Desired State (Phase 1 Findings)
  - Sample scenarios
  - Characteristics
  - Architectural Analysis
- Plans (Phase 2 Approach)
- **Conclusion**

Conclusion

- End-to-End Knowledge-Building Systems (KBSs) are needed for maximizing utilization of NASA data from missions of future in applications to benefit society
- Intelligent Archives are an essential part of such KBSs
- We have formulated a few ideas and concepts to provide recommendations that we hope will lead to
  - research by the computer science community in the near-term
  - prototyping to demonstrate feasibility in the mid-term, and
  - operational implementation in the period from 2012 to 2025.
- See http://daac.gsfc.nasa.gov/IDA/presentations.shtml for more detailed documentation

Back Up Charts

*Reports and "Drill Down" White Papers**

- Ramapriyan, H. K., Gail McConaughy, Christopher Lynnes, Steve Kempler, Ken McDonald Bob Harberts, Larry Roelofs and Paul Baker, August, 2002. *Conceptual Study of Intelligent Archives of the Future*
- Clausen, Mark and Christopher Lynnes, July, 2003. *Virtual Data Products in an Intelligent Archive*
- Harberts, Robert, L. Roelofs, H. K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald and S. Kempler, 2003. *Intelligent Archive Visionary Use Case: Advanced Weather Forecast Scenario*
- Harberts, Robert, L. Roelofs, H. K. Ramapriyan, G. McConaughy, C. Lynnes, K. McDonald and S. Kempler, September, 2003. *Intelligent Archive Visionary Use Case: Precision Agriculture Scenario*

*** (see http://daac.gsfc.nasa.gov/IDA/presentations.shtml)**

*Reports and "Drill Down" White Papers**

- Isaac, David and Christopher Lynnes, January, 2003. *Automated Data Quality Assessment in the Intelligent Archive*
- Lynnes, Christopher, July, 2003. *Automated Data Discovery and Usage*
- McConaughy, Gail and Kenneth McDonald, September, 2003. *Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges*
- Morse, H. Stephen, David Isaac, and Christopher Lynnes, January, 2003. *Optimizing Performance in Intelligent Archives*

*** (see http://daac.gsfc.nasa.gov/IDA/presentations.shtml)**

"Spin-offs"

- IS/IDU Mission Infusion activities
  - Bayesian Classification in Data Management (collaboration between GSFC and ARC) - Chris Lynnes and Kevin Wheeler
  - Wildfire Detection and Prediction (collaboration between GSFC and NOAA) – Jerry Miller et al

What Is An Intelligent Archive (IA)?

- An IA includes all items stored to support “end-to-end” research and applications scenarios
- Stored items include:
  - Data, information and knowledge (*see next chart*)
  - Software and processing needed to manage holdings and improve self-knowledge (e.g., data-mining to create robust content-based metadata)
  - Interfaces to algorithms and physical resources to support acquisition of data and their transformation into information and knowledge
- Architecture expected to be highly distributed so that it can easily adapt to include new elements as data and service providers
- Will have evolved functions beyond that of a traditional archive
- Will be based on and exploit technologies in the 10 to 20 year time range
- Will be highly adaptable so as to meet the evolving needs of science research and applications in terms of data, information and knowledge

Data, Information and Knowledge

Data: an assemblage of measurements and observations, particularly from sensors or instruments, with little or no interpretation applied

- Examples: *Scientific instrument measurements, market past performance*

Information: a summarization, abstraction or transformation of data into a more readily interpretable form

- Examples: *results after performing transformations by data mining, segmentation, classification, etc., such as a Landsat scene spatially indexed based on content, assigned a "class" value, fused with other data types, and subset for an application, for example a GIS.*

Knowledge: a summarization, abstraction or transformation of information that allows our understanding of the physical world

- Examples: *predictions from model forward runs, published papers, output of heuristics, or other techniques applied to information to answer a "what if" question such as "What will the accident rate be if an ice storm hits the Washington D.C. Beltway between Chevy Chase and the Potomac crossing at 7 a.m.?"*

*Assessment of Current State**

There has been significant progress over the last several years

- EOSDIS and ESIPs' metrics are evidence of this
- Contributions of all members of "value chain" are essential to this progress

Timeliness

- Most processes, starting from data downlink to archiving of standard products, are automated – still need operator intervention when things go wrong
- Near real-time data are sent directly to operational applications (e.g., MODIS and AIRS to NOAA; direct broadcast MODIS data used in fire monitoring)
- NOAA handling satellite data routinely to produce operational weather forecasts

Access

- Many interfaces exist for locating data from distributed set of data providers (e.g., GCMD, EOS Data Gateway, DODS/OPeNDAP, Alexandria Digital Library, Data and Information Access Link (DIAL))
- DAACs' and ESIPs' specialized access mechanisms for focused user communities
- EOS Clearing House (ECHO) - Framework available for independent development of new, specialized interfaces to locate data and services

* In addressing "what's needed"

*Assessment of Current State**

Understandability

- ATBDs, Guide documents
- Quality summaries and metadata
- Education and outreach
- Applications Program
- Education and Application REASoN Projects (*E* and *A* of REASoN)

Readiness for use

- Multiple formats are in use
- EOS data products use HDF EOS
- Tools supporting HDF EOS available
- Format conversion tools available
- Exchange of user-developed tools with each other has been useful
- Subsetting tools available for some of data, but more needed
- Not all products integrate readily into GIS

* In addressing “what’s needed”

*Assessment of Current State**

Responsiveness

- User feedback mechanisms exist for influencing system direction
- Smaller, focused systems can respond faster
- Tending towards heterogeneous and distributed systems
- Processes for defining interfaces and standards and other "framework" activities have been defined (see http://eos.nasa.gov/seeds)
- Feedback does not exist throughout the end-to-end system (e.g., an applications user cannot request sensors to collect additional data based on analysis results)

"Effortless" transformation from data to information to knowledge

- Not available today and needs attention for the future

* In addressing "what's needed"

Autonomy in an IA

- Holdings' Management Autonomy
  - Provides data to a science knowledge base in the context of research activities
  - Can exploit and use collected data in the context of a science enterprise
  - Is aware of its data and knowledge holdings and is constantly searching new and existing data for unidentified objects, features or processes
  - Facilitates derivation of information and knowledge using algorithms for Intelligent Data Understanding
  - Works autonomously to identify and characterize objects and events, thus enriching the collections of data, information and knowledge
- User Services Autonomy
  - Recognizes the value of its results, indexes/formats them properly, and delivers them to concerned individuals
  - Interacts with users in human language and visual imagery that can be easily understood by both people and machines

Autonomy in an IA

- System Management Autonomy
  - Works with other autonomous information system functions to support research
  - Manages its resources, activities and functions from sensor to user
  - Is aware of and manages the optimization of its own configuration
  - Observes its own operation and improves its own performance from sensors to models
  - Has awareness of the “state” of its cooperating external partners